Vol.2019-MUS-123 No.23
Vol.2019-SLP-127 No.23
2019/6/22

Bidirectional Gated Recurrent Units を用いた歌声合成に関する検討

片平 健太[1]　足立 優司[2]　田井 清登[2]　高島 遼一[1]　滝口 哲也[1]

概要：本研究では，オペラ歌唱音声を対象とする深層学習を用いた統計的パラメトリック歌声合成に関して，データ量と合成音声の品質の関係の検討を行う．楽譜から歌声音声を合成するシステムは近年の深層学習を用いた手法により合成音声の品質が急速に向上している．これらのシステムで合成される歌声音声は一般的な歌声音声が多いが，本研究ではより表現豊かな音声としてオペラ歌唱音声を対象とし，学習データ量と合成音声の品質の関係性を比較した．

Singing Voice Synthesis Using Bidirectional Gated Recurrent Units

1. はじめに

歌声合成システムは，任意に与えられた楽譜の歌詞情報や音高，音価などの表現から歌声を合成するシステムである．このシステムは現在主に娯楽分野において広く普及しつつある．しかし，故人の歌声の再現や病気等で声を失った患者の歌声を再現するなどの手法としても利用が可能である．

現在主に研究されている歌声合成手法として，波形接続歌声合成と統計的パラメトリック歌声合成が挙げられる．波形接続音声合成 [1] は音声波形を組み合わせることにより歌声合成する手法である．自然な音声を生成することが可能であるが，そのためには多量の合成元音声が必要となりモデルが肥大化する問題が存在する．また歌声データベースの中に滑らかに接続できる音声単位が存在しない場合には不連続を伴う合成音となり品質が劣化する．これに対し統計的パラメトリック歌声合成は，歌声データベースから統計モデルを構築し，音声パラメータを生成する．構築されたモデルは前者に対して比較的小さいものとなる．これまで主に隠れマルコフモデル (HMM) を用いた手法 [2] が広く研究されてきたが，近年ではより高品質な歌声を合成できる深層学習を用いた統計的パラメトリック歌声合成手法 [3], [4] が提案されている．

[1] 神戸大学
Kobe University
[2] メック株式会社
MEC Company Ltd.

これらの歌声合成の対象音声は一般的な歌唱音声を用いてモデル学習を行うが，本研究ではより特殊性，専門性の高い歌唱音声であるオペラ歌唱音声を用いる．オペラ歌唱では一般的な歌唱と比較して母音の発音や周波数分布が異なり，聴者の聴こえ易さに影響している．これらの要因が歌声のオペラらしさを特徴づけている．この一般的な歌唱とオペラ歌唱の相違点を上手く捉えることは，オペラという枠組みを越え音声合成や声質変換などに応用が可能である．例として雑音状況下においても聞き取り易いスピーチ音声の生成や音声に抑揚などの表現を付加することで感情を表現したり，説得力のある音声を生成することなどが考えられる．

深層学習において，学習データの量と生成データの品質には強い相関が見られ，歌声合成あるいはテキスト音声合成 (Text to Speech: TTS) 分野においてもその影響は確認されている [5]．本研究ではこのオペラ歌唱音声を目標音声とした歌声合成において，学習データ量が合成品質やオペラらしさに与える影響について検討する．

2. オペラ歌唱音声の合成

2.1 オペラ歌唱音声

一般的な歌唱音声とオペラ歌唱音声には様々な違いが存在する．一般歌唱音声とオペラ歌唱音声のスペクトログラムをそれぞれ**図 1**，**図 2** に示す．

一般歌唱音声とオペラ歌唱音声のスペクトログラムを比較すると，オペラ歌唱では 3000Hz から 4000Hz の中高

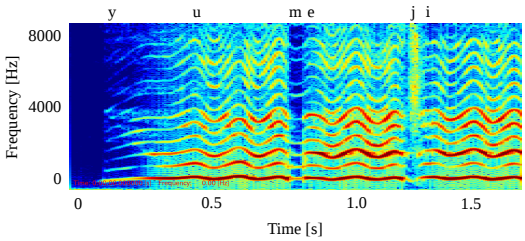

図 1　オペラ歌唱音声のスペクトログラム
Fig. 1　Spectrogram of opera singing voice.

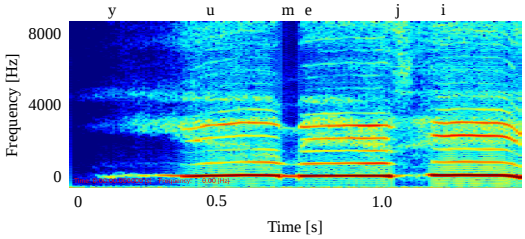

図 2　一般歌唱音声のスペクトログラム
Fig. 2　Spectrogram of singing voice.

音域の周波数成分が多く含まれる．この帯域は歌声の第 3 フォルマントから第 4 フォルマントに対応し，歌声の響き，艶，聴こえ易さに影響を与える．通常の歌唱音声では 450kHz 付近が周波数成分のピークであるが，この帯域はオペラ歌唱におけるオーケストラの伴奏の周波数成分が最も多い帯域でもあり，周波数成分が重複して歌声が伴奏に埋もれる可能性がある．そのためオペラ歌唱ではオーケストラ伴奏の周波数のピークと重複しない中高音域の周波数成分を強調することで，オーケストラの伴奏の中でも聴衆に歌声と演奏の聞き分けを容易にさせている [6]．

また，オペラ歌唱は母音の発声時に周波数成分の多い帯域が上下に移動するのが分かる．これは音程の揺らぎであるビブラートであり歌声を装飾する．オペラ歌唱では一般歌唱に比べてビブラートが深く，顕著に現れることが確認できる．

2.2　歌声合成

歌声合成システムは楽譜情報から歌声音声を合成するシステムであり，録音した歌声音声を切り取って音声を作成する波形接続型と，確率的な音響モデリングを行う統計的歌声合成法が存在する．また統計的な歌声合成法として，HMM や深層学習を用いたものが挙げられる．

深層学習による統計的パラメトリック歌声合成は，歌詞付きの楽譜情報をディープニューラルネットワーク (DNN) によって構成されるモデルに入力し，歌声音声波形を出力する．一般的には楽譜特徴量を入力とし音響特徴量を出力とする音響モデルと音響特徴量を入力とし音声波形を出力するボコーダモデルによって構成される．

2.3　楽譜特徴量

楽譜特徴量は，楽譜中の歌詞を音素単位で解析し得られる特徴量である．主な特徴として以下のものが含まれる．

- 当該音素と前後 2 つまでの音素
- 音節中における音素の位置
- 音節内の音素数
- 音符中における音節の位置
- 音符の音程，継続長，フレーズ・小節中の位置
- 音符のタイ，スラーの有無
- 音符の強弱，クレシェンド・デクレシェンド中の位置
- 総フレーズ・小節・音節・音素数
- ...

楽譜特徴量を DNN の入力として使用する場合は，これらの特徴量データをバイナリや連続値，one-hot vector で表現したものを用いる．

2.4　音響特徴量

音響特徴量は音声波形データからボコーダを用いた解析によって得られるフレーム単位の特徴量である．ボコーダから直接得られる特徴量は声色を表すスペクトル，音程を表す基本周波数，声のかすれを表す非周期成分であるが，実際に音響特徴量として使われるものはこれらを変換して次元圧縮したメルケプストラム，対数基本周波数，帯域非周期成分であることが多い．

ボコーダから得られる音響特徴量は静的特徴量であり，ここから前後の時間での値との差分である動的特徴量を求めることができる．静的特徴量を $\boldsymbol{c}$，フレーム番号を t とすると，1 次動的特徴量 $\Delta^{(1)}$ と，2 次動的特徴量 $\Delta^{(2)}$ は以下の式で求められる．

$$\Delta^{(1)}\boldsymbol{c}_t = \Delta\boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}) \quad (1)$$

$$\Delta^{(2)}\boldsymbol{c}_t = \Delta\Delta\boldsymbol{c}_t = \boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1} \quad (2)$$

2.5　Gated Recurrent Units を用いた音響特徴量推定

本研究では音響モデルに関して，オペラ歌唱音声を DNN 歌声合成の手法を用いて生成する．本研究で用いた音響モデルの学習の流れを図 **3** に示す．楽譜データから抽出された楽譜特徴量は Bidirectional GRU Network に入力され，トラジェクトリ学習により静的音響特徴量系列を得る．その後，静的音響特徴量系列から系列内変動を求め，それぞれを教師データと比較し，求められる誤差を用いてネットワークの重みを更新する．

合成時は音響モデルに楽譜特徴量を入力して得られる音響特徴量からボコーダを用いて音声を合成する．

2.5.1　Bidirectional Gated Recurrent Units

学習時，音響モデルの入力として楽譜特徴量，教師としてオペラ歌唱音声から抽出された音響特徴量を用いる．楽譜

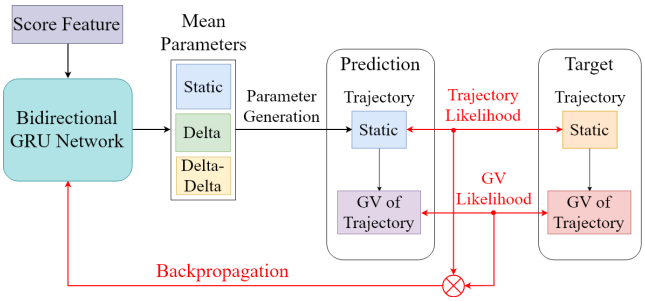

図 3　音響モデルの学習の流れ
Fig. 3　Training flow for acoustic models.

特徴量は Gated Recurrent Unit (GRU)[7] からなるネットワークに入力される．GRU は長期間の記憶を行うユニットであり，パラメータ数が少なく，かつ Long Short-Term Memory (LSTM) と同等の性能を示す特徴がある．また Bidirectional GRU を用いることで，過去の変動と未来の変動を考慮した学習を行うことができる．

2.5.2　トラジェクトリ学習

GRU ネットワークの出力 $\boldsymbol{o}$ は音響特徴の静的特徴量と 2 次までの動的特徴量がフレーム単位で結合したものである．

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\intercal \quad \Delta^{(1)}\boldsymbol{c}_t^\intercal \quad \Delta^{(2)}\boldsymbol{c}_t^\intercal\right]^\intercal \tag{3}$$

$$\boldsymbol{o} = \left[\boldsymbol{o}_1^\intercal \quad \boldsymbol{o}_2^\intercal \quad \cdots \quad \boldsymbol{o}_T^\intercal\right]^\intercal \tag{4}$$

ここから尤もらしい静的特徴量を取り出す手法として MLPG アルゴリズム [8] が存在するが，本手法ではモデルの一部に MLPG アルゴリズムを組み込んで学習するトラジェクトリ学習 [9] を行った．ここで，$\boldsymbol{\lambda}$ はパラメータである．

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{o} \mid \boldsymbol{\lambda}) = \arg\max_{\boldsymbol{c}} P(\boldsymbol{W}\boldsymbol{c} \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \tag{5}$$

$\boldsymbol{c} = [\boldsymbol{c}_1^\intercal, \cdots, \boldsymbol{c}_T^\intercal]^\intercal$ とすると，式 (3), (4) より $\boldsymbol{o}$ は $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$ のように表される一意の窓行列 $\boldsymbol{W}$ が存在する．

またここで $\hat{\boldsymbol{\mu}}$ をネットワークの出力とし，$\hat{\boldsymbol{\Sigma}}$ は別途推定した共分散行列とする．これらを用いて音響特徴量のトラジェクトリ $\hat{\boldsymbol{c}}$ を求めることが可能である．なお，$\boldsymbol{P}$ はフレーム間相関を表す．

$$\hat{\boldsymbol{c}} = \boldsymbol{P}\boldsymbol{W}^\intercal\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}} = \boldsymbol{P}\boldsymbol{W}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{o} \tag{6}$$

$$\boldsymbol{P} = (\boldsymbol{W}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{W})^{-1} \tag{7}$$

その後得られたトラジェクトリと教師データの静的特徴量との尤度 $\mathcal{L}_{trj}$ を求める．音響モデルはこの尤度を最大化するよう学習され，ネットワークの重みが更新される．なお，Z は正規化項である．

$$\mathcal{L}_{trj} = \frac{1}{Z}P(\boldsymbol{o} \mid \boldsymbol{\lambda}) = P(\boldsymbol{c} \mid \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c} \mid \hat{\boldsymbol{c}}, \boldsymbol{P}) \tag{8}$$

2.5.3　系列内変動

DNN による学習が進むと生成データの平滑化が起こる．特にメルケプストラムにおいては周波数の高帯域では変動が見られなくなる．これを抑制するため，学習時の評価関数に音響特徴量の系列内変動 (GV) に対するペナルティ [9] を設ける．GV $\boldsymbol{v}$ は時系列における静的特徴量ベクトルの分散であり，D 次元の静的特徴量 $\boldsymbol{c}$ では以下の式のように表される．

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(c_t(d) - \bar{c}(d))^2 \tag{9}$$

$$\bar{c}(d) = \frac{1}{T}\sum_{t=1}^{T}c_t(d) \tag{10}$$

ここで $\boldsymbol{v}(\boldsymbol{c}) = [v(1), \cdots, v(D)]^\intercal$ とすると，GV に対する尤度 $\mathcal{L}_{gv}$ は以下の式となる．

$$\mathcal{L}_{gv} = P(\boldsymbol{v}(\boldsymbol{c}) \mid \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\boldsymbol{v}}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{c}) \mid \boldsymbol{v}(\hat{\boldsymbol{c}}), \boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{11}$$

音響モデルの評価関数 $\mathcal{L}$ は GV 尤度に対する重み係数 w を用いると，トラジェクトリ尤度 $\mathcal{L}_{trj}$ と GV 尤度 $\mathcal{L}_{gv}$ を w で重みづけたものを掛け合わした形で表現される．

$$\mathcal{L} = \mathcal{L}_{trj}{\mathcal{L}_{gv}}^{wT} \tag{12}$$

2.5.4　音高正規化学習

DNN は統計的パラメトリック学習であるため，学習データ数が少ない音高情報が入力された場合，基本周波数の推定に失敗する可能性がある．これに対処するため実際の周波数と楽譜上の音程の差分を学習データとして使用する音高正規化学習 [10] を行う．歌声合成時には推定された基本周波数の差分と楽譜からの音符の音高を加算したものを歌声の基本周波数とする．なお，休符の場合は休符直前，直後の周波数で線形補間を行いその差分を学習に用いる．合成時には，推定後に休符に該当する箇所の周波数を 0Hz にすることで休符を復元する．

3.　実験評価

3.1　実験条件

実験には，2 つの総曲数の異なるデータセットを用いて比較を行った．**song29** は女性オペラ歌手 1 名について歌唱音声 29 曲からなる約 45 分の音声データセットである．また，**song48** は **song29** のデータに同じオペラ歌手による歌唱音声 19 曲を追加した総曲数 48 曲，約 93 分の音声データセットである．**song29** では 29 曲中 26 曲を音響モデルの学習に，3 曲をテストに用いた．**song48** では 48 曲中 43 曲を音響モデルの学習に，5 曲をテストに用いた．

いずれのデータセットも音声のサンプリング周波数は 16 kHz, フレーム長は 256 サンプルとした．

楽譜特徴量はそれぞれの歌唱音声に対応する MusicXML 形式で記述された楽譜データから抽出した 534 次元のデータを用いた．音響モデルで推定する音響特徴量には，WORLD[11] によって抽出されるスペクトル包絡から計算したメルケプストラム 59 次元，対数基本周波数 1 次元，帯域非周期成分 1 次元とこれらの 2 次までの動的特徴量に加えて有声・無声パラメータ 1 次元を用いた．楽譜特徴量，音響特徴量は共に平均 0，分散 1 になるよう事前に正規化を行った．

音響モデルは 2.5 節で述べたものを用いた．GRU ネットワークの構造は 1024 ユニットを 3 層重ねたものとした．音響モデルの学習は，**song29**, **song48** ともに GRU ネットワークのみを学習し (gru)，そのネットワークを使用してトラジェクトリ学習を行い (trj)，最後に GV を考慮した学習を行う (gvtrj) 3 段階に分けて行った．以降，**song29** は総曲数 29 での gvtrj まで行ったモデルとする．入力データのバッチサイズは gru は 2，trj, gvtrj は 1 である．gvtrj における GV 尤度の重みは $w = 1.0 \times 10^{-6}$ とした．深層学習を用いたモデルでは Adam[12] による誤差逆伝播法を用いて学習を行った．Adam のパラメータは初期学習率 1.0×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$ とした．歌声波形はモデルによって推測される音響特徴量を WORLD による合成を行うことで得られる．

本研究では主観評価として平均オピニオン指標 (MOS) を用いたオペラ性の測定を行った．これは初めにプロのオペラ歌唱音声と一般的な歌唱音声を聴き，その後合成音声を聴いて 1 を一般歌唱，5 をオペラ歌唱として，合成音声がどちらに近いか 5 段階評価を行ったものである．被験者は 11 人でテストデータからランダムに抽出された 20 フレーズに対して評価を行った．また客観評価としてメルケプストラム歪み (MCD)，基本周波数二乗平均平方根誤差 (F_0 RMSE)，有声・無声パラメータ偽陽性率　(V/UV FPR)，偽陰性率　(V/UV FNR)，帯域非周期性成分歪み (BAPD)，系列内変動距離 (GVD) を用いた．いずれも数値が低いほど精度は高くなる．

3.2　実験結果と考察

図 4 に **song29** と **song48** をそれぞれ用いて学習した音響モデルで生成した歌声のオペラ性に関する MOS 評価の結果を示す．なお，gru, trj, gvtrj は **song48** の各学習終了時の生成結果である．図のエラーバーは 95%信頼区間を表す．

学習データ量の違いについて着目すると，**song29** と **song48** の gvtrj は有意な差が見られ，**song48** を用いて学習したモデルが **song29** を用いて学習したモデルよりも評価は高いことが分かる．また，**song48** の学習の段階について比較すると，段階が進むほどオペラ性の評価は向上している．

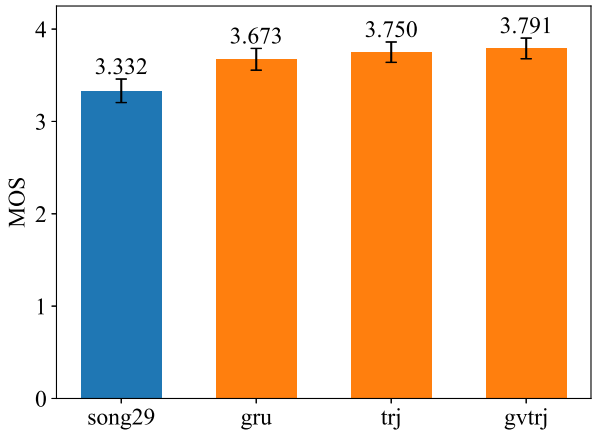

図 4　オペラらしさに関する MOS 評価
Fig. 4　MOS evaluation for opera-style.

表 1 に **song29** と **song48** をそれぞれ用いて学習した音響モデルで生成した歌声の客観評価を示す．**song48** の gvtrj は MCD, F_0 RMSE, BAPD, GVD において **song29** よりも優れた値を示した．**song48** での各学習での結果を比較すると，GVD において gvtrj はそれ以前のモデルでの数値より低い値を示しており，GV に対するペナルティの効果が見られた．一方で，MCD など一部の指標では数値の上昇がみられる．これは全ての音響特徴量に対して同時に推定を行っているため，一部の特徴量に対して過学習が起きている可能性がある．

図 5 は **song29** での合成音声, **song48** での合成音声 (gvtrj), 目標音声の F_0 の軌跡である．この歌声音声では開始から 1 秒付近までは D#4 の音程で発声しており，gvtrj の軌跡は **song29** のものより目標音声に近いことが分かる．また，この音程は **song29** では 90 回出現し，**song48** では 182 回出現しており，学習データの増加が F_0 推定の改善に影響を与えたと考えられる．

図 6, **図 7**, **図 8** は **song29** での合成音声, **song48** での合成音声 (gvtrj), 目標音声のスペクトログラムである．合成音声にもオペラの歌唱音声の特徴である中高音域のフォルマントの強調やビブラートが確認できる．また，gvtrj のスペクトルは **song29** のものと比較して目標音声のより細かな変化を捉えており，これがオペラ性の評価の向上に繋がったと考えられる．

これらより，学習データ量が増加するとオペラ性が向上することが分かる．また，トラジェクトリ学習や GV を考慮した学習が合成音声の品質を改善させることが分かった．

4.　おわりに

本稿では，学習データ量と GRU ネットワークによる合

表 1　客観評価
Table 1　Result of objective evaluations.

	MCD (dB)	F_0 RMSE (cent)	V/UV FPR	V/UV FNR	BAPD (dB)	GVD
song29	5.267	58.66	6.245×10^{-2}	7.208×10^{-3}	17.76	8.679×10^{-2}
gru	5.058	57.96	6.699×10^{-2}	8.591×10^{-3}	16.88	9.228×10^{-2}
trj	5.153	58.72	7.033×10^{-2}	8.059×10^{-3}	17.07	1.878×10^{-1}
gvtrj	5.180	58.49	7.092×10^{-2}	7.880×10^{-3}	17.17	8.049×10^{-2}

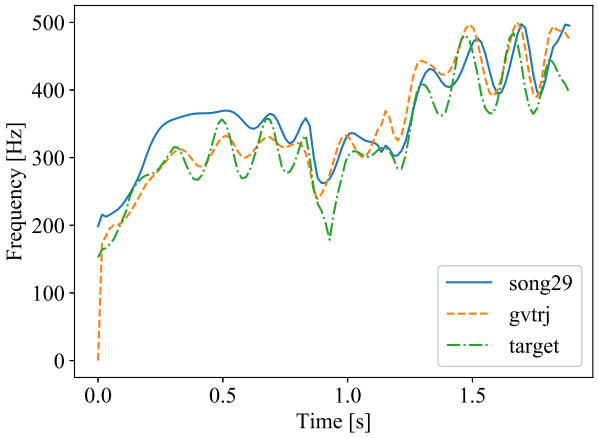

図 5　F_0 の比較
Fig. 5　Comparison of F_0.

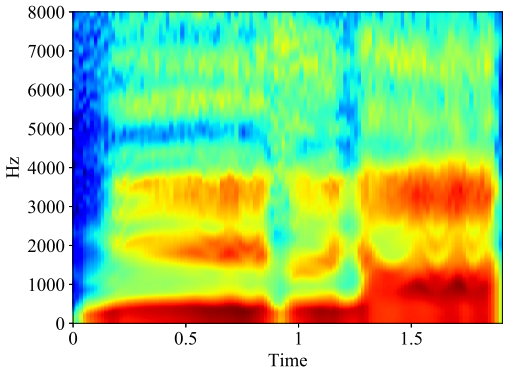

図 6　**song29** の合成音声のスペクトログラム
Fig. 6　Spectrogram of **song29**.

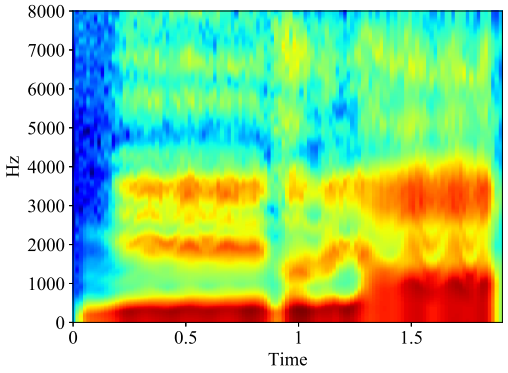

図 7　**song48**(gvtrj) の合成音声のスペクトログラム
Fig. 7　Spectrogram of **song48** (gvtrj).

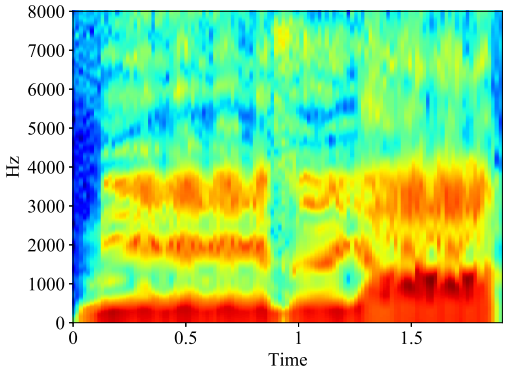

図 8　目標音声のスペクトログラム
Fig. 8　Target spectrogram.

成歌声音声の品質の関係について検討した．図 4 より学習データ量が多いほど音響モデルが生成する合成音声のオペラ性の観点から見た品質は向上した．

本研究では合成音声の品質を高める手段としてデータの量の観点から問題に取り組んだが，今後はデータの質の観点から品質を高める手法を検討する．例として一般歌唱音声とオペラ歌唱音声の音素の発音の違いを音素ラベルなどの入力特徴量に付加する手法などが挙げられる．

音響特徴量の推定に関して，現在はすべての特徴量を 1 つのモデルで同時に推定しているが，各特徴の最適モデルはそれぞれ異なることから，音響モデルを特徴ごとに分割して学習を行うことでより高精度な推定を行える可能性がある．

また，近年ではボコーダモデルに WaveNet[13] など深層学習を用いたものが提案されており，これらを用いてより高品質な音声を生成することを検討する．

参考文献

[1] Bonada, J., Umbert, M. and Blaauw, M.: Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016., *Proc. Interspeech*, pp. 1230–1234 (2016).
[2] Saino, K. et al.: An HMM-based singing voice synthesis system, *Ninth International Conference on Spoken Language Processing* (2006).
[3] 法野行哉ほか: Deep Neural Network に基づく歌声合成システム - Sinsy，日本音響学会 秋季研究発表会 講演論

文集，pp. 1099–1102 (2018).
[4] 片平健太ほか：深層学習を用いた歌声合成の検討，日本音響学会 春季研究発表会講演論文集，pp. 1091–1092 (2019).
[5] 林知樹ほか：WaveNet ボコーダにおける学習データ量の影響に関する調査，日本音響学会春季研究発表会 講演論文集，pp. 249–250 (2018).
[6] Sundberg, J. et al.: 歌声の科学，pp. 122–123（オンライン），入手先 ⟨http://ci.nii.ac.jp/ncid/BA81510991⟩, 東京電機大学出版局 (2007).
[7] Chung, J. et al.: Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).
[8] Tokuda, K. et al.: Speech parameter generation algorithms for HMM-based speech synthesis, *ICASSP*, Vol. 3, pp. 1315–1318 (2000).
[9] Hashimoto, K. et al.: Trajectory training considering global variance for speech synthesis based on neural networks, *ICASSP*, pp. 5600–5604 (2016).
[10] Nishimura, M. et al.: Singing Voice Synthesis Based on Deep Neural Networks, *Proc. Interspeech*, pp. 2478–2482 (2016).
[11] Morise, M. et al.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).
[12] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
[13] Van Den Oord, A. et al.: WaveNet: A generative model for raw audio, *SSW*, p. 125 (2016).